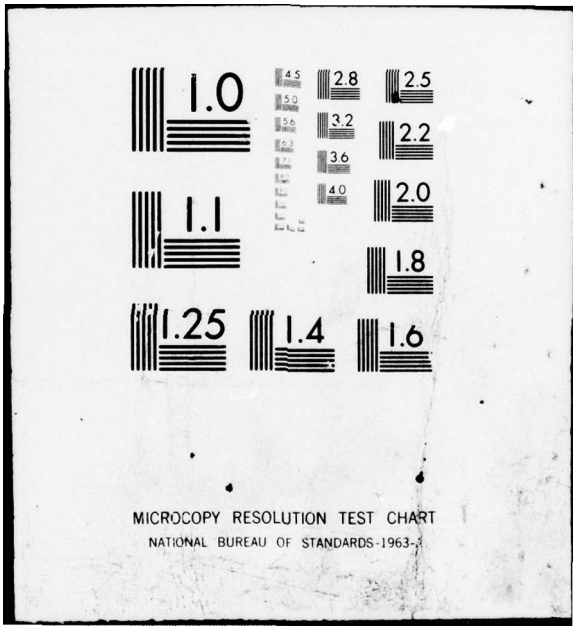
AD-A079 733 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
FURTHER STUDY OF ROBUSTIFICATION VIA A BAYESIAN APPROACH.(U)
SEP 79 G CHEN, G E BOX DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1998 NL

END
DATE FILMED
2-80
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MRC Technical Summary Report # 1998

FURTHER STUDY OF ROBUSTIFICATION VIA A BAYESIAN APPROACH

Gina Chen and George E. P. Box

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

September 1979

(Received July 27, 1979)

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Traingle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

FURTHER STUDY OF ROBUSTIFICATION VIA A BAYESIAN APPROACH

Gina Chen and George E. P. Box

Technical Summary Report # 1998
September 1979

ABSTRACT

The Bayesian outlier procedure discussed by Box and Tiao (1968) which uses the contaminated normal model is further explored in this report. For a simple location estimate suggested by their method, the weight given each observation is expressed explicitly in terms of standardized residuals so allowing comparison with M-estimators.

AMS (MOS) Subject Classification: 62G35

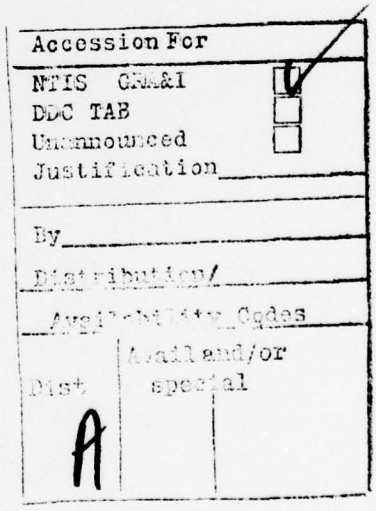
Key Words: Contaminated normal model, Bayesian procedure, M-estimator, weighting function, standardized residuals

Work Unit Number 4: Probability, Statistics, and Combinatorics

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

SIGNIFICANCE AND EXPLANATION

A Bayesian model has been proposed which describes the generation of an observation by a process whereby with prior probability $1-\alpha$ the usually assumed statistical structure is correct but with small probability α it is incorrect (for example, the observation has a very large variance). For a simple location estimate the nature of the down weighting of outlying observations produced by this model is studied and is compared with that of the presently popular M-estimators.



FURTHER STUDY OF ROBUSTIFICATION VIA A BAYESIAN APPROACH

Gina Chen and George E. P. Box

1. Introduction

Statistics is a science which helps extract meaningful information from data which are subject to errors. Two important aspects of statistical analysis are

(i) The building of probabilistic models, defining the functional form of a relationship between variables as well as the distribution of the errors.

(ii) The development of techniques of analysis which are efficient supposing the data to be derived from the specified model.

Model building begins with preliminary and graphic analysis of the data. This information is combined with theoretical knowledge about the system to produce a tentative model. One desirable criterion for choosing a model is simplicity. Such a model is easy to interpret and can give greater precision than less parsimonious models (Box, 1978). An iterative model building procedure is needed which can move from a perhaps initially faulty model to one which is adequate for the purpose at hand.

Two techniques are of value in this process.

(i) Robustification: By employing robust methods (or models) we are unlikely to be misled by those model inadequacies guarded against.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

(ii) Iterative fixing: Conditional inference can be made assuming that the tentative model is true, and one can then make diagnostic checks to reveal possible inadequacies. If there are indications of inadequacy, the model can be modified appropriately.

Both techniques have their advantages and limitations. It is impossible to make a procedure robust against every possible contingency. On the other hand, situations can arise where tests may fail to show up inadequacies that could, nevertheless, cause serious problems. Looking for inadequacies after the event may not always be successful.

A practical policy is to robustify for one or more contingencies that seem likely in the context of a particular application and then to study the residuals to seek for the unexpected.

The Standard Statistical Models

Consider the standard statistical model

$$y_i = \eta(\underset{\sim}{x}_i, \underset{\sim}{\theta}) + \varepsilon_i \qquad i = 1,\ldots,n \qquad (1.1)$$

where ε_i's are independently, identically and normally distributed. In this model, it is assumed that

(i) The observation y_i is a function of dependent variables $\underset{\sim}{x}_i$ and parameter $\underset{\sim}{\theta}$ plus an error ε_i.

(ii) The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independently distributed.

(iii) All the errors ε_i's have the same normal distribution with a fixed variance.

Violations of any of these assumptions can cause problems. However, this thesis is concerned only with departures from the third assumption.

It is well known that real data can be affected with discordant values produced by a mistake or a sudden change of some unknown factor. It is usually, therefore, unrealistic to assume that every observation is from the same normal distribution. A more sensible possibility suggested by Dixon (1953) and by Tukey (1960) is to allow errors to be from two different distributions. The standard distribution $f(\varepsilon)$ occurs with probability $1-\alpha$ and an alternative distribution $g(\varepsilon)$ generates discordant values with probability α. We will call this the contaminated model. From, for example, the study made in Chapter 3 of this thesis, it is clear that this is not the only possibility. As we saw there, a model which could be realistic in some cases would have errors generated from one distribution up to a certain point and then randomly switching to a different distribution having either a different mean or a different variance or both for a certain period of time and then switching again. A model which

has this property for switches in means was proposed by George Barnard (1959) in his justification of the cumulative sum chart. We will only consider the contaminated model.

2. A Bayesian Analysis of the Contaminated Model

In those situations where a contaminated distribution is appropriate, a point estimate may be obtained by taking the mean of the posterior distribution of the appropriate parameter and Bayesian intervals can be calculated. The posterior mean takes the form of a weighted average in which the weights are posterior probabilities. The use of the Bayesian posterior mean as an estimator may be justified on the grounds that it minimizes the mean square error loss.

The formulation of Box and Tiao (1968) can be readily generalized to include nonlinear models as follows.

Consider a model $Y = \eta(X,\theta) + \varepsilon$ where Y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of fixed elements with rank p, θ is a $q \times 1$ vector of parameters, $\eta(X,\theta)$ is an $n \times 1$ vector for each given X and θ, and ε is an $n \times 1$ vector of random errors. Suppose now that each of the errors independently comes from either one of the two distributions — a standard distribution $f(\varepsilon|\xi_1)$ and an alternative distribution $g(\varepsilon|\xi_2)$, where ξ_1 and ξ_2 are nuisance parameters and θ is the parameter of interest.

Let $a_{(r)}$ be the event that a particular set of r of the n ε's are from $g(\varepsilon|\xi_2)$ and the remaining $s = n-r$ from $f(\varepsilon|\xi_1)$. Corresponding to $a_{(r)}$, vector ε was partitioned into $\varepsilon_{(r)}$, $\varepsilon_{(s)}$; vector Y into $Y_{(r)}$, $Y_{(s)}$ and matrix X into $X_{(r)}$, $X_{(s)}$. Let $P^{(r)}$ be the prior probability of event $a_{(r)}$, $P(\theta,\xi_1,\xi_2)$ be the prior distribution of the parameters. The posterior distribution is then given by

$$P(\theta|Y) = \sum_{(r)} P(a_{(r)}|Y)P(\theta|a_{(r)},Y)$$

where $P(\theta|a_{(r)},Y) =$

$$\frac{\int P(\theta,\xi_1,\xi_2)\dot{f}(Y_{(s)}-\eta(X_{(s)},\theta)/\xi_1)\dot{g}(Y_{(r)}-\eta(X_{(r)},\theta)/\xi_2)d\xi_1 d\xi_2}{\int P(\theta,\xi_1,\xi_2)\dot{f}(Y_{(s)}-\eta(X_{(s)},\theta)/\xi_1)\dot{g}(Y_{(r)}-\eta(X_{(r)},\theta)/\xi_2)d\xi_1 d\xi_2 d\theta}$$

where $\dot{f}$ denotes the product of densities of the elements of $Y_{(s)}-\eta(X_{(s)},\theta)$ and $\dot{g}$ that of $Y_{(r)}-\eta(X_{(r)},\theta)$. The denominator is the probability of the event that the $\varepsilon_{(r)}$ are drawn from $g(\varepsilon|\xi_2)$ and the $\varepsilon_{(s)}$ from $f(\varepsilon|\xi_1)$. If we denote it by $h(Y_{(r)}\sim g, Y_{(s)}\sim f)$ then

$$P(a_{(r)}|Y) = \frac{P^{(r)}h(X_{(r)}\sim g; Y_{(s)}\sim f)}{\sum_{(r)} P^{(r)}h(Y_{(r)}\sim g; Y_{(s)}\sim f)} = c\,\frac{P^{(r)}h(Y_{(r)}\sim g; Y_{(s)}\sim f)}{P^{(o)}h(Y_{(r)}\sim f; Y_{(s)}\sim f)}$$

$$= c\,\frac{P^{(r)}h(Y_{(s)}\sim f)h(Y_{(r)}\sim g|Y_{(s)}\sim f)}{P^{(o)}h(Y_{(s)}\sim f)h(Y_{(r)}\sim f|Y_{(s)}\sim f)} = c\,\frac{P^{(r)}h(Y_{(r)}\sim g|Y_{(s)}\sim f)}{P^{(o)}h(Y_{(r)}\sim f|Y_{(s)}\sim f)}$$

where $h(Y_{(s)} \sim f)$ is the marginal probability that $\varepsilon_{(s)}$ is from $f(\varepsilon|\xi_1)$, $h(Y_{(r)} \sim g|Y_{(s)} \sim f)$ is the conditional probability that $\varepsilon_{(r)}$ is from $g(\varepsilon|\xi_2)$ given that $\varepsilon_{(s)}$ is from $f(\varepsilon|\xi_1)$.

3. The Contaminated Normal Distribution

Since all experimental data are subject to unpredictable mistakes, a probability model which would be more realistic than the standard normal model would often be one where with probability $1-\alpha$, the error is normally distributed with mean 0 and variance σ^2 and with small probability α the error has a mean 0 but an inflated variance $k^2\sigma^2$.

One argument in favor of this model is as follows. The trimmed mean has been recommended as a sensible estimator for location parameters by Tukey and McLaughlin (1963), Huber (1972) and Stigler (1977). More recently M-estimators have been recommended by Huber, Hampel and Andrews (Andrews et al (1972)). From the study in TSR# 1997 we have seen that the contaminated normal distributions would produce via the Bayesian route estimators which are very similar to those recommended by these authors.

Furthermore, results of TSR# 2002 suggest that heavy-tailed distributions are sometimes caused by inhomogeneity in mean and variance such as might be encountered early in an experiment because of start-up difficulties. After such inhomogeneity has been allowed for, the observations seem to be adequately repre-

sented by a contaminated normal distribution. Thus, for a carefully planned experiment, a contaminated normal distribution is likely to be appropriate.

4. Bayesian Inference with Contaminated Normal Model

Consider that the error distribution is a contaminated normal distribution $(1-\alpha)N(0,\sigma^2) + \alpha N(0,k^2\sigma^2)$. This is the special case where $f(\varepsilon|\xi_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\varepsilon^2}{2\sigma^2}}$, $g(\varepsilon|\xi_2) = \frac{1}{\sqrt{2\pi}k\sigma} e^{-\frac{\varepsilon^2}{2k^2\sigma^2}}$ and $P^{(r)} = \alpha^r(1-\alpha)^{n-r}$. Assume, a priori that the information about θ and about σ are independent and, as in Box and Tiao (1968) approximate the locally noninformative prior by $P(\theta,\sigma) \propto P(\theta)\cdot\frac{1}{\sigma}$. Then

$$P(\theta,\xi_1,\xi_2) = P(\theta,\sigma) \propto P(\theta)\,\frac{1}{\sigma}$$

and if we let $w_{(r)} = P(a_{(r)}|Y)$ and $P_{(r)}(\theta|Y) = P(\theta|a_{(r)},Y)$ we have

$$P(\theta|Y) = \sum_{(r)} w_{(r)}P_{(r)}(\theta|Y)$$

with

$$P_{(r)}(\theta|Y) \propto \int_0^\infty \sigma^{-(n+1)}P(\theta)\exp\left\{-\frac{1}{2\sigma^2}S_{(r)}(\theta)\right\}d\sigma$$

$$\propto P(\theta)[S_{(r)}(\theta)]^{-\frac{n}{2}}$$

where $S_{(r)}(\theta) = (Y_{(s)}-\eta(X_{(s)},\theta))'(Y_{(s)}-\eta(X_{(s)},\theta))$

$$+\frac{1}{k^2}(Y_{(r)}-\eta(X_{(r)},\theta))'(Y_{(r)}-\eta(X_{(r)},\theta))$$

and

$$w_{(r)} = c\,\frac{p^{(r)}h(Y_{(r)}\sim g; Y_{(s)}\sim f)}{p^{(o)}h(Y_{(r)}\sim f; Y_{(s)}\sim f)}$$

$$= c\left(\frac{\alpha}{1-\alpha}\right)^{r} \frac{\int\sigma^{-(n+1)}p(\theta)k^{-r}e^{-\frac{1}{2}S_{(r)}(\theta)\sigma^{-2}}d\theta d\sigma}{\int\sigma^{-(n+1)}p(\theta)e^{-\frac{1}{2}[(Y-\eta(X,\theta))'(Y-\eta(X,\theta))]\sigma^{-2}}d\theta d\sigma}$$

$$= c\left(\frac{\alpha}{1-\alpha}\right)^{r}k^{-r}\,\frac{\int P(\theta)[S_{(r)}(\theta)]^{-\frac{n}{2}}d\theta}{\int P(\theta)[(Y-\eta(X,\theta))'(Y-\eta(X,\theta))]^{-n/2}d\theta}$$

where c is some constant.

When η is approximately linear in θ over the range considered, it would also be reasonable to take $P(\theta)$ as approximately constant. In such case the above formula becomes

$$P_{(r)}(\theta|Y) \propto \{S_{(r)}(\theta)\}^{-\frac{1}{2}n}$$

$$w_{(r)} = c\left(\frac{\alpha}{1-\alpha}\right)^{r}k^{-r}\,\frac{\int\{S_{(r)}(\theta)\}^{-\frac{n}{2}}d\theta}{\int\{(Y-\eta(X,\theta))'(Y-\eta(X,\theta))\}^{-n/2}d\theta}$$

A robust point estimate for θ is then given by the posterior mean

$$\int \theta P(\theta|Y)d\theta = \sum_{(r)} w_{(r)}\int\theta P_{(r)}(\theta|Y)d\theta\ .$$

Furthermore, probabilities can be calculated which are useful in spotting possible bad values. (Observations generated by g are referred to as bad values.) If we let $w_{(r)}$ be the posterior probability of event $a_{(r)}$ given the sample Y:

a_i be the event that the observation y_i is from g and all the others are from f.

a_{ij} be the event that the observations y_i and y_j are from g and all the others are from f.

Then w_i, w_{ij} are the posterior probabilities associated with a_i and a_{ij}.

Using these individual probabilities, one can calculate

q_1, the posterior probability that only one observation is from g is $\sum_{i=1}^{n} w_i$

q_2, the posterior probability that only two observations are from g is $\sum_{\substack{i,j=1 \\ i<j}}^{n} w_{ij}$

and similarly q_3, q_4,...,etc.

Also, the posterior conditional probabilities can be calculated as follows:

$q_{i/1}$ = P(the i-th observation is from g/ there is only one bad value) = w_i/q_1

$q_{ij/2}$ = P(the i-th and j-th observations are from g/there are only two bad values) = w_{ij}/q_2

.
.
.

etc.

The distribution g is regarded as the source of "bad values," and these probabilities can be interpreted as the posterior probabilities that there is one bad value, two bad values,...,etc., and the conditional probabilities as the posterior probabilities that the i_1th, i_2th,..., i_ℓth observations are bad values given that there are ℓ bad values.

Although the posterior probabilities q_i's depend on the choices of α and k, the conditional probabilities are independent of α and rather insensitive to k. It seems that, in practice by comparing these conditional probabilities one can readily identify questionable observations.

5. Some Further Study of the Linear Model

When $\eta(X,\theta) = X\theta$, We have the general linear model and the posterior distribution of θ can be written as in Box and Tiao in the following way

$$P(\theta|Y) = \sum_{(r)} w_{(r)} P_{(r)}(\theta|Y)$$

with

$$w_{(r)} = c\left(\frac{\alpha}{1-\alpha}\right)^r k^{-r} \frac{|X'X|^{1/2}}{|X'X-\phi X_{(r)}'X_{(r)}|^{1/2}} \left\{\frac{s^2_{(r)}}{s^2}\right\}^{-\frac{1}{2}\nu} \quad (5.1)$$

$$P_{(r)}(\theta|Y) = \frac{\Gamma(\frac{1}{2}n)|X'X-\phi X_{(r)}'X_{(r)}|^{1/2}}{\Gamma(\frac{1}{2}\nu)(\pi\nu s^2_{(r)})^{\frac{1}{2}p}} \times \left\{1 + \frac{(\theta-\hat{\theta}_{(r)})(X'X-\phi X_{(r)}'X_{(r)})(\theta-\hat{\theta}_{(r)})}{\nu s^2_{(r)}}\right\}^{-\frac{1}{2}n} \quad (5.2)$$

where

$$\hat{\theta}_{(r)} = (X'_{(n-r)}X_{(n-r)} + \frac{1}{k^2}X'_{(r)}X_{(r)})^{-1}(X'_{(n-r)}Y_{(n-r)} + \frac{1}{k^2}X'_{(r)}Y_{(r)}) \quad (5.3)$$

$$s^2_{(r)} = \frac{1}{\nu}S_{(r)}(\hat{\theta}_{(r)}) = \frac{1}{\nu}\Big\{(Y_{(n-r)}-X_{(n-r)}\hat{\theta}_{(r)})' \times (Y_{(n-r)}-X_{(n-r)}\hat{\theta}_{(r)}) + \frac{1}{k^2}(Y_{(r)}-X_{(r)}\hat{\theta}_{(r)})'(Y_{(r)}-X_{(r)}\hat{\theta}_{(r)})\Big\} \quad (5.4)$$

$\nu = n-p$ and $s^2 = s^2_{(o)}$.

The posterior distribution $P_{(r)}(\theta|Y)$ is a p-dimensional multivariate t-distribution with mean $\hat{\theta}_{(r)}$, dispersion matrix $s^2_{(r)}(X'X-\phi X'_{(r)}X_{(r)})^{-1}$ *and* $\nu = n-p$ *degrees of freedom.* It is then easy to see that the robust point estimate - the posterior mean - is equal to $\sum_{(r)} w_{(r)}\hat{\theta}_{(r)}$. As mentioned before, the posterior mean is justified because it minimizes the mean square error loss.

The Weighting Structure of this Bayesian Posterior Mean

A most important feature of a robust estimator is the weighting pattern it puts on the observations. As expected, for an estimator to be insensitive to outlying observations, the weights given to extreme observations must be small. Different criteria can be used. In L-estimators, the weight given to an observation depends on the

percentage point it takes in the ordered sample, while in M-estimators, the weights are determined by the relative distances of the observations. It is, therefore, informative to study the weighting structure of the Bayesian posterior mean. Two things of special interest would be the variables through which the weight for each observation is determined and the roles α and k play in this estimator.

As stated previously, the posterior mean is $\sum_{(r)} w_{(r)}\hat{\theta}_{(r)}$ with formula for $w_{(r)}$ and $\hat{\theta}_{(r)}$ given by (5.1) and (5.3). Box and Tiao (1968) give the following two equations

$$\nu s^2_{(r)} = \nu s^2 - \phi(Y_{(r)}-X_{(r)}\hat{\theta})'\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}(Y_{(r)}-X_{(r)}\hat{\theta}) \tag{5.5}$$

$$\hat{\theta}_{(r)} = \hat{\theta} - \phi(X'X)^{-1}X'_{(r)}\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}(Y_{(r)}-X_{(r)}\hat{\theta}). \tag{5.6}$$

Using equation (5.6)

$$\begin{aligned} Y_{(r)}-X_{(r)}\hat{\theta}_{(r)} &= Y_{(r)}-X_{(r)}\hat{\theta}+\phi X_{(r)}(X'X)^{-1}X'_{(r)}\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1} \\ &\quad \times (Y_{(r)}-X_{(r)}\hat{\theta}) \\ &= \{I+\phi X_{(r)}(X'X)^{-1}X'_{(r)}\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}\} \\ &\quad \times (Y_{(r)}-X_{(r)}\hat{\theta}) \quad . \end{aligned} \tag{5.7}$$

Since

$$\{I+\phi X_{(r)}(X'X)^{-1}X'_{(r)}\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}\} \cdot \{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}$$

$$= I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}+\phi X_{(r)}(X'X)^{-1}X'_{(r)}$$

$$= I$$

it follows that

$$\{I+\phi X_{(r)}(X'X)^{-1}X'_{(r)}\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}\} = \{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}.$$

Equation (5.7) becomes

$$Y_{(r)}-X_{(r)}\hat{\theta}_{(r)} = \{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}(Y_{(r)}-X_{(r)}\hat{\theta})$$

so $$Y_{(r)}-X_{(r)}\hat{\theta} = \{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}(Y_{(r)}-X_{(r)}\hat{\theta}_{(r)}). \quad (5.8)$$

Dividing both sides of equation (5.5) by $\nu s^2_{(r)}$ and using equation (5.8), we have

$$\frac{s^2}{s^2_{(r)}} = 1+\frac{\phi}{\nu s^2_{(r)}}(Y_{(r)}-X_{(r)}\hat{\theta})'\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}(Y_{(r)}-X_{(r)}\hat{\theta})$$

$$= 1+\frac{\phi}{\nu s^2_{(r)}}(Y_{(r)}-X_{(r)}\hat{\theta}_{(r)})'\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}' \cdot \{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}^{-1}\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}(Y_{(r)}-X_{(r)}\hat{\theta}_{(r)})$$

$$= 1+\frac{\phi}{\nu}\left(\frac{Y_{(r)}-X_{(r)}\hat{\theta}_{(r)}}{s_{(r)}}\right)'\{I-\phi X_{(r)}(X'X)^{-1}X'_{(r)}\}\left(\frac{Y_{(r)}-X_{(r)}\hat{\theta}_{(r)}}{s_{(r)}}\right). \quad (5.9)$$

Note that $\hat{\theta}_{(r)}$ and $s^2_{(r)}$ are the estimates obtained by downweighting the observations which under the assumption that $a_{(r)}$ occurs, come from a distribution with the larger variance. Also $r_{(r)} = \frac{Y_{(r)} - X_{(r)}\hat{\theta}_{(r)}}{s_{(r)}}$ is the residual vector for the observations which are downweighted under such an assumption.

From (5.1), we can write

$$w_{(r)} = c\left(\frac{\alpha}{1-\alpha}\right)^r k^{-r} \frac{|X'X|^{\frac{1}{2}}}{|X'X - \phi X'_{(r)}X_{(r)}|^{1/2}} \left\{\frac{s^2}{s^2_{(r)}}\right\}^{\frac{1}{2}\nu}$$

$$= c\left(\frac{\alpha}{1-\alpha}\right)^r k^{-r} \frac{|X'X|^{\frac{1}{2}}}{|X'X - \phi X'_{(r)}X_{(r)}|^{1/2}} \{1 + \frac{\phi}{\nu} r'_{(r)}$$

$$\cdot \{I - \phi X_{(r)}(X'X)^{-1}X'_{(r)}\}r_{(r)}\}^{\frac{1}{2}\nu}. \qquad (5.10)$$

If we let $\Sigma = \phi(I - \phi X_{(r)}(X'X)^{-1}X'_{(r)})\frac{\nu-r}{\nu}$, then

$$|X'X|^{-1}|\Sigma| = |X'X|^{-1}|I - \phi X_{(r)}(X'X)^{-1}X_{(r)}|\left(\frac{\nu-r}{\nu}\phi\right)^r$$

$$= \begin{vmatrix} X'X & X'_{(r)} \\ \phi X_{(r)} & I \end{vmatrix} \left(\frac{\nu-r}{\nu}\phi\right)^r = \begin{vmatrix} I & \phi X_{(r)} \\ X'_{(r)} & X'X \end{vmatrix} \left(\frac{\nu-r}{\nu}\phi\right)^r$$

$$= |I||X'X - X'_{(r)}\phi X_{(r)}|\left(\frac{\nu-r}{\nu}\phi\right)^r = |X'X - \phi X'_{(r)}X_{(r)}|\left(\frac{\nu-r}{\nu}\phi\right)^r.$$

Thus

$$|\Sigma| = |X'X||X'X-\phi X'_{(r)}X_{(r)}|\left(\frac{\nu-r}{\nu}\phi\right)^{r}$$

and

$$w_{(r)} = c|X'X|\left(\frac{\alpha}{1-\alpha}\right)^{r}k^{-r}\left(\frac{\nu-r}{\nu}\phi\right)^{\frac{r}{2}}\frac{1}{|\Sigma|^{1/2}}\left(1+\frac{1}{\nu-r}r'_{(r)}\Sigma r_{(r)}\right)^{\frac{1}{2}\nu}. \tag{5.11}$$

This $w_{(r)}$ is proportional to the inverse of a multivariate-t ordinate at $r_{(r)}$ with precision matrix Σ and $\nu-r$ degrees of freedom. It is seen then that $r_{(r)}$ is an important factor which determines the weighting structure. If the components of $r_{(r)}$, which are the standardized residuals for the observations down-weighted in the process of estimation, are very large, then $w_{(r)}$ will be large and $\hat{\theta}_{(r)}$ will be a more influential constituent of the posterior mean. This is; in some sense, similar to an M-estimator where the weighting pattern is determined through the standardized residuals. However, the estimates must be calculated iteratively in that case.

6. A Method of Comparing the Weights with Those of the M-estimators

Recall that an M-estimate is the solution to the following equation

$$\sum_{i=1}^{n} \underset{\sim}{x}_i'\left(\frac{y_i-\underset{\sim}{x}_i\theta}{s}\right)w\left(\frac{y_i-\underset{\sim}{x}_i\theta}{s}\right) = 0. \tag{6.1}$$

If we write $v_i = 1/w(\frac{y_i - x_i\theta}{s})$, then (6.1) can be written as

$$\sum_{i=1}^{n} x_i' y_i v_i^{-1} - \sum_{i=1}^{n} x_i' x_i \theta v_i^{-1} = 0$$

or equivalently

$$X'V^{-1}Y - X'V^{-1}X\theta = 0$$

where

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad V = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & v_n \end{pmatrix}.$$

The solution $\hat{\theta}$ is then given by $(X'V^{-1}X)^{-1}X'V^{-1}Y$. This is the same solution as the weighted least square estimates in a linear model when $Var(Y) = V\sigma^2 = \begin{bmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_n \end{bmatrix}\sigma^2$ and the weighting function w corresponds to the weights v_i^{-1} used in the weighted least square method. In order to compare the weighting pattern for the Bayesian posterior mean with the w function in an M-estimator, one needs to write the posterior mean approximately in the form of $(X'V^{-1}X)^{-1}X'V^{-1}Y$ and the matrix V will then provide the information about the weighting structure. It is not always possible to obtain such a matrix V for the general linear model, and even when possible, its form is often complicated. However, in the special case

of a location parameter, simple results can be obtained and much insight gained.

The Location Case

Consider the model $y_i = \theta + \varepsilon_i$

ε_i i.i.d. from $(1-\alpha)N(0,\sigma^2)+ \alpha N(0,k^2\sigma^2)$, $i=1,\ldots,n$.

Then $X'X = n$, $X'_{(r)}X_{(r)} = i$, $I-\phi X_{(r)}(X'X)^{-1}X'_{(r)} = I- \frac{\phi}{n} X_{(r)}X'_{(r)}$

and following (4.5.11) $w_{(r)}$ reduces to

$$c\left(\frac{\alpha}{1-\alpha}\right)^r k^{-r}\left(\frac{n}{n-r\phi}\right)^{\frac{1}{2}}\{1+ \frac{\phi}{n-1} r'_{(r)}\{I- \frac{\phi}{n} X_{(r)}X'_{(r)}\}r_{(r)}\}^{\frac{1}{2}(n-1)} . \quad (6.2)$$

Also from (5.6)

$$\hat{\theta}_{(r)} = \bar{y} - \frac{r\phi}{n-r\phi} (\bar{y}_{(r)}-\bar{y})$$

where $\bar{y}$ is the sample average and $\bar{y}_{(r)}$ is the average of elements of $Y_{(r)}$.

The posterior mean is then

$$\sum_{(r)} c\left(\frac{\alpha}{1-\alpha}\right)^r k^{-r}\left(\frac{n}{n-r\phi}\right)^{\frac{1}{2}}\{1+ \frac{\phi}{n-1} r'_{(r)}(I- \frac{\phi}{n}X_{(r)}X'_{(r)})r_{(r)}\}^{\frac{1}{2}(n-1)}$$

$$\times (\bar{y}- \frac{r\phi}{n-r\phi}(\bar{y}_{(r)}-\bar{y})) . \quad (6.3)$$

We shall now discuss this formula in the following cases:
(i) when there is at most one discordant observation in the sample,
(ii) when there are at most two discordant observations in the sample.

Only One Discordant Observation Present

In this circumstance, the $w_{(r)}$ associated with two or more observations coming from the distribution with the larger variance will be negligible. The terms left in (6.3) would then be associated with no bad values and with one bad value.

$$c+\sum_{i=1}^{n} c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}\left\{1+\frac{\phi}{n-1}\,r_i'\left(1-\frac{\phi}{n}\right)r_i\right\}^{\frac{1}{2}(n-1)}\left(\bar{y}-\frac{\phi}{n-\phi}(y_i-\bar{y})\right)$$

$$= c+\sum_{i=1}^{n} c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}\left(1+\frac{\phi}{n-1}\,\frac{n-\phi}{n}\,r_i^2\right)^{\frac{1}{2}(n-1)}\left(\bar{y}-\frac{\phi}{n-\phi}(y_i-\bar{y})\right) \qquad (6.4)$$

where c is a constant such that the weights would sum up to unity.

Let $R_i = \left(1+\frac{\phi}{n-1}\,\frac{n-\phi}{n}\,r_i^2\right)^{\frac{1}{2}(n-1)}$ and $R = \sum_{i=1}^{n} R_i$

then $$c + \sum_{i=1}^{n} c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R_i = 1$$

or equivalently

$$c + c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R = 1$$

which implies

$$c = 1/\left(1 + \frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R\right) .$$

Substituting c, R_i and R, (6.4) becomes

$$\sum_{i=1}^{n}\left[c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R_i\,\frac{n}{n-\phi}\,\bar{y}-c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R_i\,\frac{\phi}{n-\phi}\,y_i\right]$$

$$=\sum_{i=1}^{n}\left[c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}\frac{1}{n-\phi}\,Ry_i-c\left(\frac{\alpha}{1-\alpha}\right)k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}\frac{1}{n-\phi}\,\phi R_i y_i\right]$$

$$=\frac{\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}\frac{1}{n-\phi}}{1+\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R}\sum_{i=1}^{n}(R-\phi R_i)y_i\ .$$

We, therefore, have

$$\text{posterior mean} \approx \frac{1}{1+\frac{\alpha}{1-\alpha}k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R}\,\bar{y}+\frac{\frac{\alpha}{1-\alpha}k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R}{1+\frac{\alpha}{1-\alpha}k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R}\sum_{i=1}^{n}\frac{R-\phi R_i}{(n-\phi)R}\,y_i\ . \tag{6.5}$$

Let $\bar{Y}_0 = \bar{y}$ = ordinary mean, $\bar{Y}_1 = \sum_{i=1}^{n}\frac{R-\phi R_i}{(n-\phi)R}\,y_i$,

and $$Q = \frac{1}{1+\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R}$$

the posterior mean $\approx Q\bar{Y}_0 + (1-Q)\bar{Y}_1$.

It is clear now that the posterior mean can be written as a convex combination of two location estimates, $\bar{Y}_0$ — which is the conditional posterior mean given that there is no outlier — and

$\bar{Y}_1$ — which is the conditional posterior mean given there is exactly one outlier. The quantity Q is the posterior probability that there are no outliers, and $1-Q$ is the posterior probability that there is exactly one outlier. The crucial quantities here are R_i's, which are the inverse of the t-ordinates of r_i's. If a particular observation y_i is abberant, the corresponding residual r_i will be large. By taking R_i as the inverse of the corresponding t-ordinate, contrasts with other R_j's will be exaggerated. Thus $\frac{R-\phi R_i}{(n-\phi)R}$ will be made relatively small. The quantity $\bar{Y}_1$ will then put little weight on y_i. Also if there is a discrepant observation, R will be large which will make Q small and the posterior mean close to $\bar{Y}_1$. Conversely, if there are no discrepant observations, the posterior mean will lie close to $\bar{Y}$.

The value of Q depends not only on the sample through R, but also on values of α and k. If one assumes a larger prior probability α of outliers, then the posterior probability of no outlier could be smaller. Notice too that the posterior mean depends on α only through Q so that the role that α plays is only to adjust the overall weight associated with fixed numbers of outliers.

When k is fairly large, $\phi = 1 - \frac{1}{k^2}$ would be close to 1 ($\phi = .96$ when $k = 5$). From the formula of R_i and (6.5), it

seems that $\bar{Y}_1$ would be rather insensitive to k since it involves only ϕ and so is R. If we let $G = \frac{\alpha}{1-\alpha} k^{-1}$ and approximate ϕ by 1, then the posterior mean involves only one constant G. One interpretation associated with G is the following.

Assume y_1 is an observation from $(1-\alpha)N(\theta,\sigma^2)+\alpha N(\theta,k^2\sigma^2)$ then G is the ratio of the posterior probability that y_1 is from $N(\theta,k^2\sigma^2)$ to the posterior probability that y_1 is from $N(\theta,\sigma^2)$ given that $y_1 = \theta$.

$$\frac{P(y_1 \sim N(\theta,k^2\sigma^2)/y_1=\theta)}{P(y_1 \sim N(\theta,\sigma^2)/y_1=\theta)} = \frac{P(y_1 \sim N(\theta,k^2\sigma^2), y_1=\theta)}{P(y_1 \sim N(\theta,\sigma^2), y_1=\theta)}$$

$$= \frac{P(y_1=\theta/y_1 \sim N(\theta,k^2\sigma^2))P(y_1 \sim N(\theta,k^2\sigma^2))}{P(y_1=\theta/y_1 \sim N(\theta,\sigma^2))P(y_1 \sim N(\theta,\sigma^2))}$$

$$\frac{\frac{1}{\sqrt{2\pi}k\sigma} \cdot \alpha}{\frac{1}{\sqrt{2\pi}\sigma}(1-\alpha)} = \frac{\alpha}{1-\alpha}\frac{1}{k} = G .$$

G can also be viewed as

$$G = \frac{\text{expected value of } \sqrt{\text{information from } N(0,k^2\sigma^2)}}{\text{expected value of } \sqrt{\text{information from } N(0,\sigma^2)}}$$

$$= \frac{\text{probability of an observation being from } N(0,k^2\sigma^2) \cdot \sqrt{\text{information in } N(0,k^2\sigma^2)}}{\text{probability of an observation being from } N(0,\sigma^2) \cdot \sqrt{\text{information in } N(0,\sigma^2)}}$$

$$= \frac{\alpha\sqrt{1/k^2\sigma^2}}{1-\alpha\sqrt{1/\sigma^2}} = \frac{\alpha}{1-\alpha}\frac{1}{k} .$$

So for k large, we need to determine only one parameter in the model, namely G.

The Weighting Pattern

In the case of a single location parameter,

$(X'V^{-1}X)^{-1}X'V^{-1}Y = \sum_{i=1}^{n} \frac{v_i^{-1}}{\sum_{1}^{n} v_i^{-1}} y_i$. Thus the weighting function

$V^{-1} = \begin{pmatrix} v_1^{-1} & & \\ & \ddots & \\ & & v_n^{-1} \end{pmatrix}$ used in obtaining the weighted least square estimate is proportional to the weight given to each observation. It is, therefore, possible to write the posterior mean in the form of $(X'V^{-1}X)^{-1}X'V^{-1}Y$ and compare V^{-1} with the weighting patterns used in the M-estimators.

If we assume at most one outlier, the posterior mean is $Q\bar{Y}_0 + (1-Q)\bar{Y}_1$ as shown above. It can also be written as $\sum_{i=1}^{n} \left(\frac{Q}{n} + (1-Q) \frac{R-\phi R_i}{(n-\phi)R} \right) y_i$. Therefore,

$$v_i^{-1} \propto \frac{Q}{n} + (1-Q) \frac{R-\phi R_i}{(n-\phi)R} .$$

The right-hand side is a function of r_i; we can thus obtain the weighting pattern of this Bayesian posterior mean in terms of the standardized residuals. This weighting pattern is sample dependent. We illustrate with a random sample of size 10 from N(0,1).

Ten observations generated from a computer random sampling routine and the corresponding residuals r_i's and weights (v_i^{-1})'s with $\alpha = .05$ and $k = 5.0$ are listed in Table 1. It is seen that the weight given to each observation is approximately constant and close to .1. The weights change when a discrepant observation is present. We have added in turn 1, 2, 3, 4, 5, to the seventh observation and again calculated the r_i and v_i^{-1}; the results are also shown in Table 1.

If we plot v_7^{-1} versus r_7, with 0, 1, 2, 3, 4, 5 added to y_7, we found that they lie on a smooth curve and such a curve is comparable with the weighting function of the M-estimators since they both determine the weights given to observations as a function of properly standardized residuals. Figure 1 gives plots of v_7^{-1} versus r_7 for $\alpha = .01, .05$ and $.10$ and $k = 5.0$. In terms of G, they are weighting curves for $G = .002, .011$ and $.022$. The weighting curves for M-estimators are shown in Figure 2. Examination of Figure 1 shows that the weights are roughly equal for residuals between zero and one and start to descend as they go farther and farther away from the center. The larger α is (or the larger G is) the sooner and faster the weights decrease. Comparing with Figure 2 this weighting pattern has the desired property (Beaton and Tukey, 1974) that it is almost as constant in

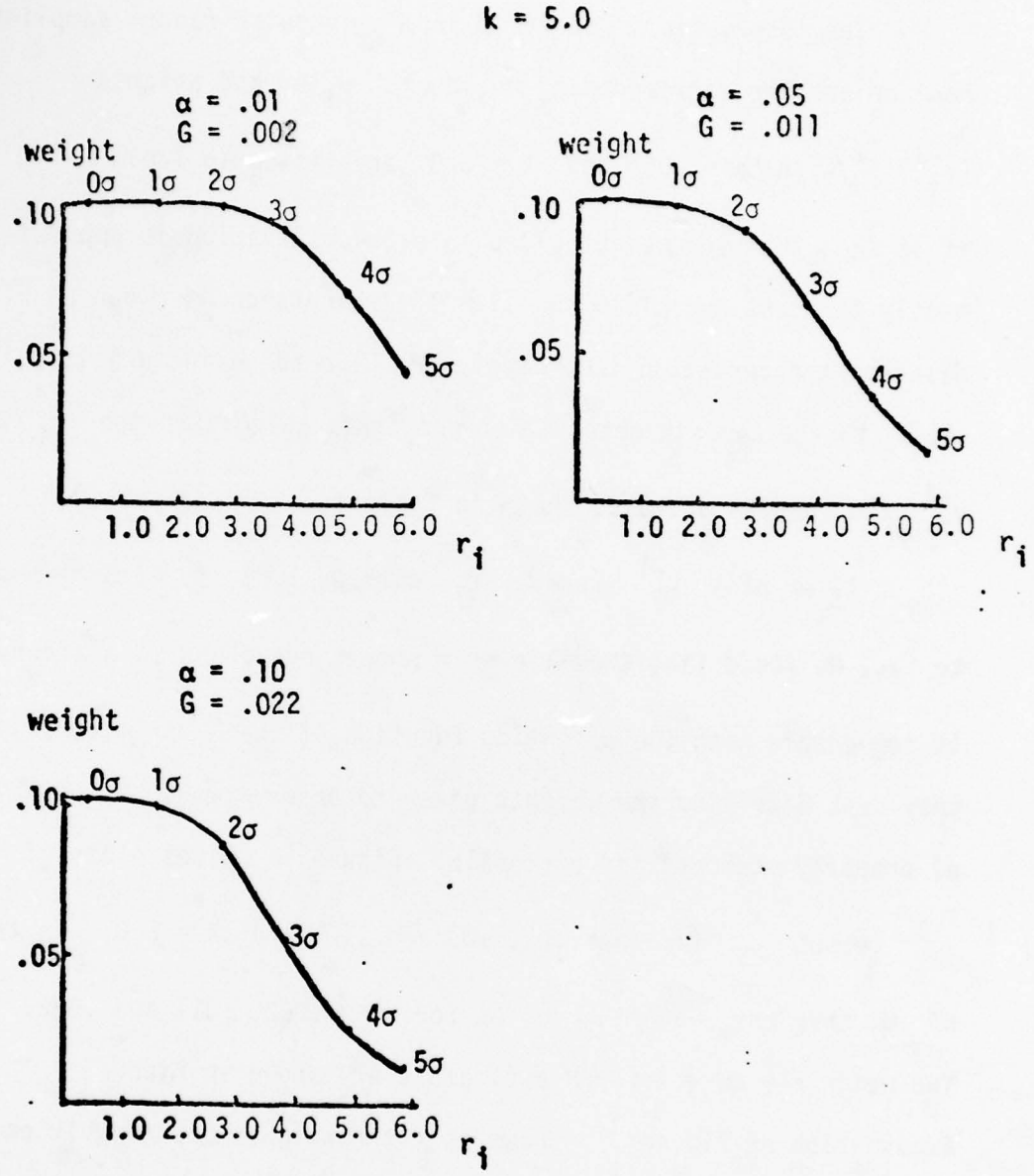


Figure 1 Weighting curves for different $\alpha(G)$ values with $k = 5.0$.

i	1	2	3	4	5	6	7	8	9	10
y_i	-.21	.23	.17	-1.24	-1.09	1.23	.52	-.18	1.16	.92
r_i	-.47	.10	.03	-2.15	-1.85	1.54	.48	-.43	1.42	1.04
v_i^{-1}	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
add 1 to 7-th observation						$y_7 = 1.52$				
r_i	-.54	-.03	-.09	-2.03	-1.77	1.21	1.65	-.50	1.11	.79
v_i^{-1}	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
add 2 to 7-th observation						$y_7 = 2.52$				
r_i	-.55	-.12	-.17	-1.75	-1.55	.88	2.78	-.52	.81	.56
v_i^{-1}	.10	.10	.10	.10	.10	.10	.09	.10	.10	.10
add 3 to 7-th observation						$y_7 = 3.52$				
r_i	-.54	-.18	-.22	-1.50	-1.35	.64	3.87	-.52	.58	.38
v_i^{-1}	.10	.10	.10	.10	.10	.10	.07	.10	.10	.10
add 4 to 7-th observation						$y_7 = 4.52$				
r_i	-.52	-.22	-.26	-1.31	-1.19	.46	4.90	-.50	.42	.25
v_i^{-1}	.11	.11	.11	.11	.11	.11	.04	.11	.11	.11
add 5 to 7-th observation						$y_7 = 5.52$				
r_i	-.50	-.25	-.28	-1.16	-1.06	.34	5.85	-.49	.30	.15
v_i^{-1}	.11	.11	.11	.11	.11	.11	.02	.11	.11	.11

Table 1 The correspondences among y_i, r_i and v_i^{-1} with $\alpha = .05$, $k = 5.0$

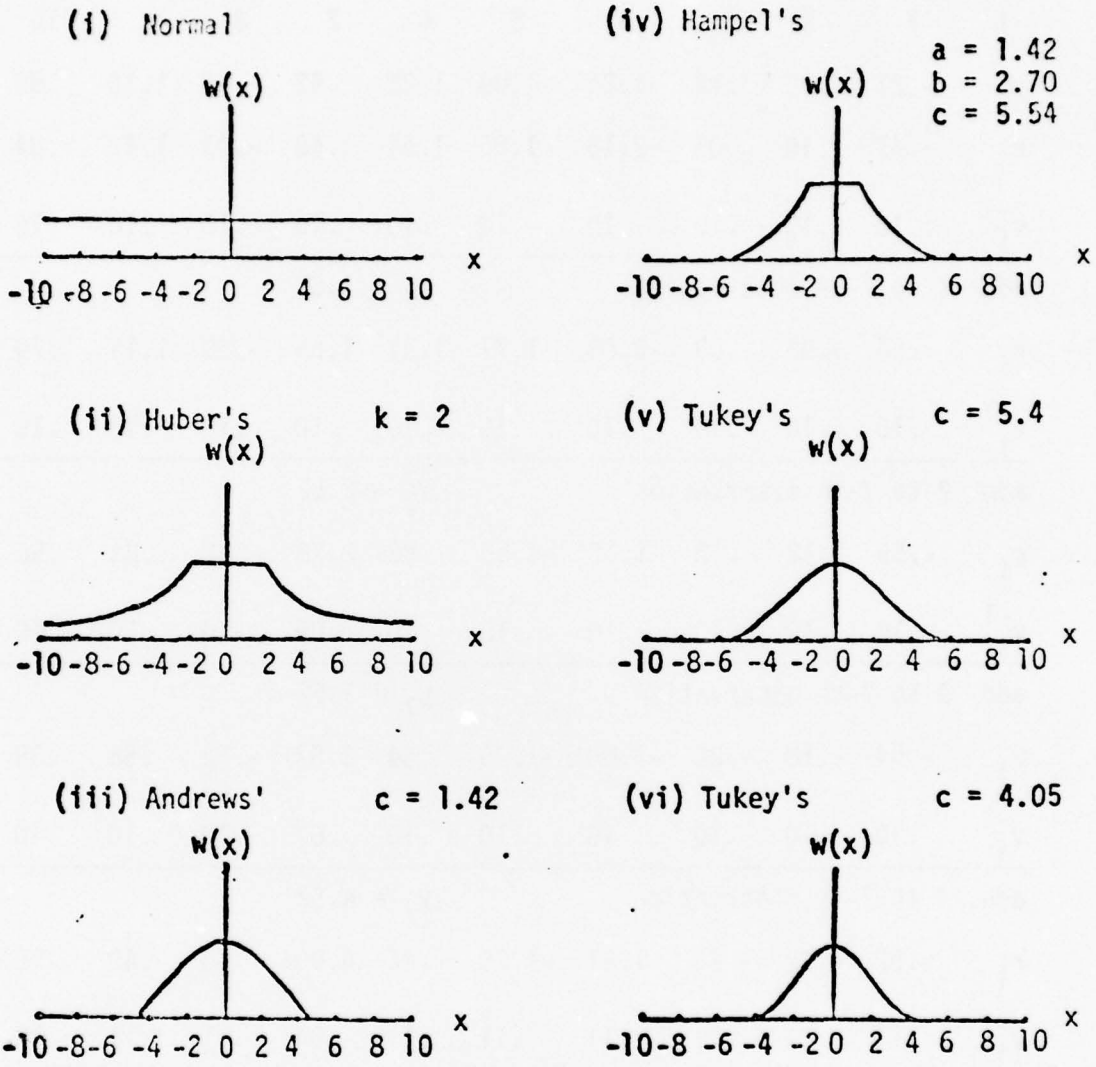


Figure 2 w-functions for various M-estimators.

the center (normal-like) as Huber's and Hampel's and reduces to small values as smoothly as Andrews' and Tukey's. Smoothness may be another desired property since a sudden change in the slope is rather artificial and one does not have reason to expect such a change.

Figure 1 depends on a particular sample. However, in all, totally five samples were drawn. All of the weighting patterns are similar and those in Figure 1 are typical.

As discussed before, when k is large, G seems to be the only relevant parameter to be determined. Figure 1 has shown the weighting patterns for $G = .002, .011$ and $.022$ where systematic changes have been observed. With this same sample, one can also fix G and calculate the weights for different α and k values. Figure 3 exhibits five weighting patterns all with $G = .011$ and the following pairs of α and k values.

α	.050	.095	.136	.174	.208
k	5.0	10.0	15.0	20.0	25.0

All five curves look very much alike with the greatest discrepancy being that for the values $(\alpha,k) = (.05, 5.0)$. This confirms our conjecture that G is the only critical parameter when $k \geq 5$.

At Most Two Discordant Observations Present

As we have seen, the terms in (6.3) corresponding to no bad values and one bad value can be written as $c\bar{y}$ and

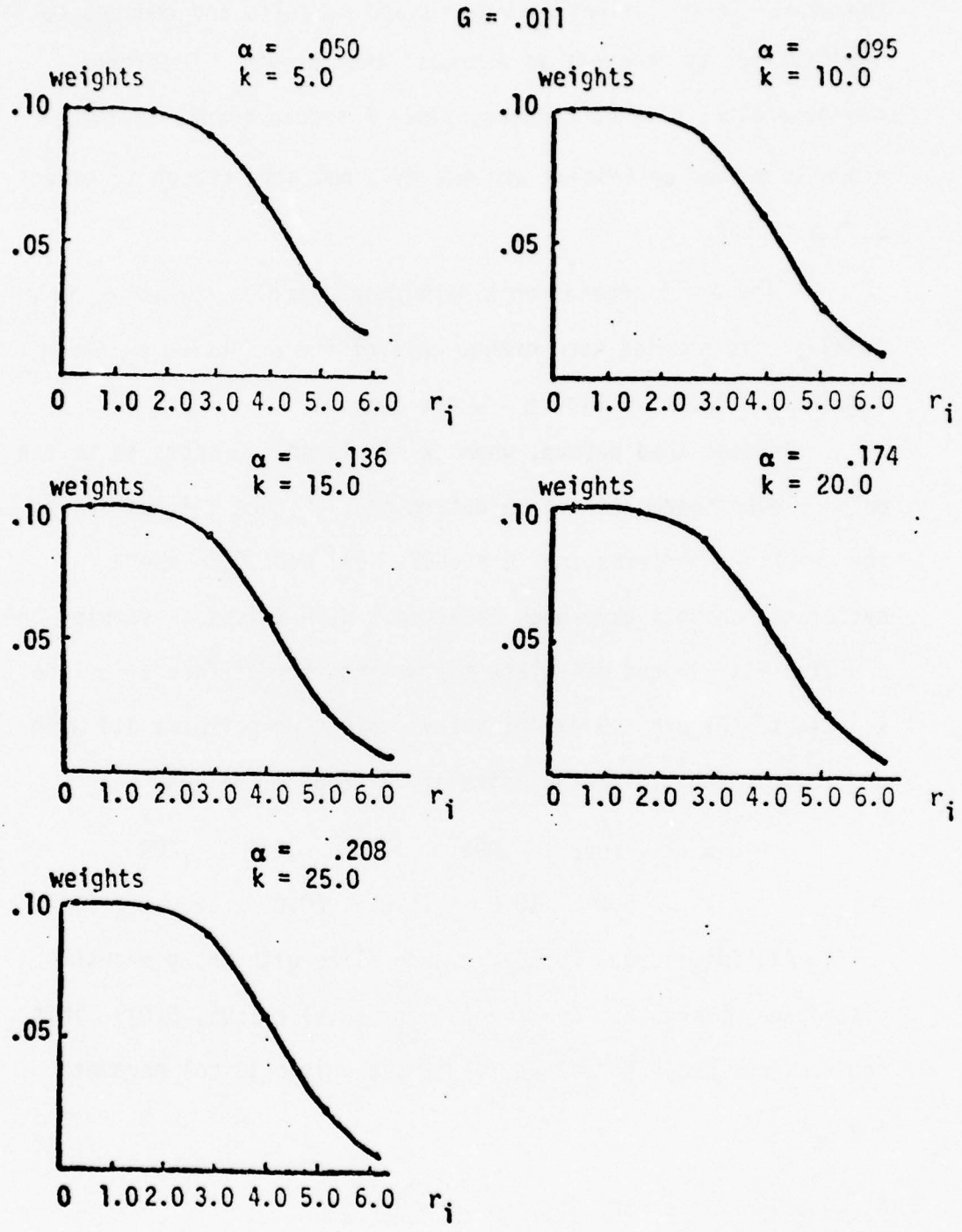


Figure 3 Weighting patterns for fixed G and different α,k values.

$c\,\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}} R \sum_{i=1}^{n} \frac{R-\phi R_i}{(n-\phi)R}\, y_i$. The term corresponding to two bad values is

$$\sum_{\substack{i,j=1 \\ i<j}}^{n} c\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} \{1+\frac{\phi}{n-1}\, r'_{ij}(I-\frac{\phi}{n}\, X'_{ij})r_{ij}\}^{\frac{1}{2}(n-1)} \times (\bar{y} - \frac{2\phi}{n-2\phi}\,(\bar{y}_{ij}-\bar{y}))$$

$$= \sum_{\substack{i,j=1 \\ i<j}}^{n} c\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} \left\{1+\frac{\phi}{n-1}\, r'_{ij}\begin{pmatrix} \frac{n-\phi}{n} & -\frac{\phi}{n} \\ -\frac{\phi}{n} & \frac{n-\phi}{n} \end{pmatrix} r_{ij}\right\}^{\frac{1}{2}(n-1)} \times \left(\frac{n}{n-2\phi}\,\bar{y} - \frac{2\phi}{n-2\phi}\,\bar{y}_{ij}\right) \tag{6.6}$$

where $r_{ij} = \begin{pmatrix} r^i_{ij} \\ r^j_{ij} \end{pmatrix}$, $r^i_{ij} = \frac{y_i - \hat{\theta}_{ij}}{s_{ij}}$, $r^j_{ij} = \frac{y_j - \hat{\theta}_{ij}}{s_{ij}}$,

$\bar{y}_{ij} = \frac{y_i + y_j}{2}$, and $\hat{\theta}_{ij}$ and s_{ij} are estimates obtained by weighted least squares in which the i-th and j-th observations are downweighted.

If we let $T_{ij} = \left\{1+\frac{\phi}{n-1}\, r'_{ij}\begin{pmatrix} \frac{n-\phi}{n} & -\frac{\phi}{n} \\ -\frac{\phi}{n} & \frac{n-\phi}{n} \end{pmatrix} r_{ij}\right\}^{\frac{1}{2}(n-1)}$

$$= \left\{1+\frac{\phi}{n-1}\left((r^i_{ij})^2+(r^j_{ij})^2 - \frac{\phi}{n}(r^i_{ij}+r^j_{ij})^2\right)\right\}^{\frac{1}{2}(n-1)} \tag{6.7}$$

and $T = \frac{1}{2} \sum_{\substack{i,j \\ i\neq j}} T_{ij}$,

(6.7) reduces to

$$\sum_{\substack{i,j=1\\ i<j}}^{n} c\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} T_{ij}\left(\frac{n}{n-2\phi}\bar{y} - \frac{2\phi}{n-2\phi}\frac{y_i+y_j}{2}\right)$$

$$= c\sum_{i=1}^{n}\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} T\,\frac{1}{n-2\phi}\,y_i - c\sum_{\substack{i,j=1\\ i<j}}^{n}\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}$$

$$\times\ \frac{\phi}{n-2\phi}\,T_{ij}(y_i+y_j)\ .$$

It is easy to see that $T_{ij} = T_{ji}$ so

$$\sum_{\substack{i,j=1\\ i<j}}^{n} T_{ij}(y_i+y_j) = \sum_{\substack{i,j=1\\ i<j}}^{n} T_{ij}y_i + \sum_{\substack{i,j=1\\ i<j}}^{n} T_{ji}y_j\ ,$$

$$= \sum_{\substack{i,j=1\\ i<j}}^{n} T_{ij}y_i + \sum_{\substack{i,j=1\\ i>j}}^{n} T_{ij}y_i = \sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} T_{ij}\right) y_i\ .$$

The above formula can now be written as

$$c\sum_{i=1}^{n}\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} T\,\frac{1}{n-2\phi}\,y_i - c\sum_{i=1}^{n}\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}\frac{\phi}{n-2\phi}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} T_{ij}\right)y_i$$

$$= c\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}\frac{1}{n-2\phi}\sum_{i=1}^{n}\left(T-\phi\sum_{\substack{j=1\\ j\neq i}}^{n} T_{ij}\right) y_i$$

$$= c\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} T\sum_{i=1}^{n}\frac{\left(T-\phi\sum_{\substack{j=1\\ j\neq i}}^{n} T_{ij}\right)}{(n-2\phi)T}\,y_i\ .$$

Thus the posterior mean will be

$$c\bar{y}+c\,\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R\sum_{i=1}^{n}\frac{R-\phi R_i}{(n-\phi)R}\,y_i+c\left(\frac{\alpha}{1-\alpha}\right)^{2}k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}T\sum_{i=1}^{n}\frac{\left(T-\phi\sum_{\substack{j=1\\ j\neq i}}^{n}T_{ij}\right)}{(n-2\phi)T}\,y_i.$$

We have

$$c+c\,\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R+c\left(\frac{\alpha}{1-\alpha}\right)^{2}k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}T=1$$

which implies

$$c=1\Big/\left(1+\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R+\left(\frac{\alpha}{1-\alpha}\right)^{2}k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}T\right).$$

If now

$$\bar{Y}_0=\bar{y}$$

$$\bar{Y}_1=\sum_{i=1}^{n}\frac{(R-R_i)}{(n-\phi)R}\,y_i$$

$$\bar{Y}_2=\sum_{i=1}^{n}\frac{\left(T-\phi\sum_{\substack{j=1\\ j\neq i}}^{n}T_{ij}\right)}{(n-2\phi)T}\,y_i\quad,$$

we can write the posterior mean as

$$\frac{1}{1+\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R+\left(\frac{\alpha}{1-\alpha}\right)^{2}k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}T}\,\bar{Y}_0$$

$$+\frac{\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}R}{1+\frac{\alpha}{1-\alpha}\,k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}}R+\left(\frac{\alpha}{1-\alpha}\right)^{2}k^{-2}\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}}T}\,\bar{Y}_1$$

$$+ \frac{\left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right) T}{1+\frac{\alpha}{1-\alpha} k^{-1}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}} R + \left(\frac{\alpha}{1-\alpha}\right)^2 k^{-2}\left(\frac{n}{n-2\phi}\right) T} \bar{Y}_2 \ .$$

Again, let $G = \frac{\alpha}{1-\alpha}\frac{1}{k}$ and

$$Q_0 = \frac{1}{1+G\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}} R+G^2\left(\frac{n}{n-2\phi}\right) T}, \quad Q_1 = \frac{G\left(\frac{n}{n-2\phi}\right)^{\frac{1}{2}} R}{1+G\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}} R+G^2\left(\frac{n}{n-2\phi}\right) T},$$

$$Q_2 = 1 - Q_0 - Q_1,$$

the posterior mean is then a weighted average of three estimators, $Q_0\bar{Y}_0 + Q_1\bar{Y}_1 + Q_2\bar{Y}_2$, where $\bar{Y}_i$ is the conditional posterior mean given that there are exactly i bad values and Q_i is the posterior probability that there are i bad values.

Generalization to the case of ℓ-bad values is not difficult but is tedious. In general, however,

(1) The posterior mean for ℓ-outliers is a weighted average $Q_0\bar{Y}_0+\ldots+Q_\ell\bar{Y}_\ell$ of $\ell+1$ estimators where $\bar{Y}_i$ is the conditional posterior mean given that there are exactly i bad values and Q_i is the posterior probability of i bad values.

(2) If k is large then ϕ can be taken to be 1 and G is the only parameter to be determined in advance in the model.

(3) Each $\bar{Y}_i$ itself is a weighted average with the weight for the j-th observation heavily dependent on the series of quantity

$$\left\{1 + \frac{\phi}{n-1} r'_{jj_1\cdots j_{i-1}} \begin{pmatrix} \frac{n-\phi}{n} & \cdots & -\frac{\phi}{n} \\ \vdots & & \vdots \\ -\frac{\phi}{n} & \cdots & \frac{n-\phi}{n} \end{pmatrix} r_{jj_1\cdots j_{i-1}}\right\}^{\frac{1}{2}(n-1)}$$

$$= \left\{1 + \frac{\phi}{n-1}\left((r^{j}_{jj_1\cdots j_{i-1}})^2 + (r^{j_1}_{jj_1\cdots j_{i-1}})^2 + \ldots + (r^{j_{i-1}}_{jj_1\cdots j_{i-1}})^2 - \frac{\phi}{n}(r^{j}_{jj_1\cdots j_{i-1}} + r^{j_1}_{jj_1\cdots j_{i-1}} + \ldots + r^{j_{i-1}}_{jj_1\cdots j_{i-1}})^2\right)\right\}^{\frac{1}{2}(n-1)} \quad (6.8)$$

for all possible combinations of $j_1, j_2, \ldots, j_{i-1}$.

In this expression, $r_{jj_1\cdots j_{i-1}}$ is the residual vector for observations $y_j, y_{j_1}, \ldots, y_{j_{i-1}}$ calculated from the estimate where these observations are down weighted. If the j-th observation itself is a bad value, then (6.8) will always be large since $r^{j}_{jj_1\cdots j_{i-1}}$ will be large. When $\{y_j, y_{j_1}, \ldots, y_{j_{i-1}}\}$ are exactly the i bad values in the sample, the above quantity reaches its maximum. This will result in a small weight for the j-th observation. If the j-th

observation is not a bad value then no matter how we choose $j_1,\ldots,j_{i-1}$, the expression (6.8) will never reach its maximum and so the weight given to the j-th observation will not be very small.

(4) Equation (6.8) is a direct extension of (6.7). From the form of (6.7), it is seen that, if the absolute values of r_{ij}^{i} and r_{ij}^{j} were fixed, the value of (6.7) would be larger when they were both positive or both negative and smaller when they had opposite signs. This implies that if there are two bad values in a sample, these bad values will have less weight if they fall on different sides of the mean and more weight if they fall on the same side. For the case of ℓ bad values, similarly, equation (6.8) indicates that the bad values will have less weight if they are equally spread on both sides of the mean and more weight if more of them are concentrated on the same side.

Example: Darwin's Data

We illustrate how observations are weighted in such a Bayesian analysis using Darwin's data concerning fifteen differences of the heights of cross- and self-fertilized plants quoted by Fisher (1960, p 37). Our interest here will be to examine more closely the weighting structure of the Bayesian posterior mean, and to further develop the analysis of Box and Tiao (1968).

The data consist of measurements on 15 pairs of plants, each pair contained a self-fertilized and a cross-fertilized plant grown in the same pot. The 15 differences y_i were recorded in Table 2 in ascending order.

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
-67	-48	6	8	14	16	23	24
y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	
28	29	41	49	56	60	75	

Table 2 Darwin's Data

Inspection of the data indicates that two observations -67 and -48 are remote from the rest. We shall assume there are at most two bad values in the sample and $k = 5$. Then the posterior mean $\hat{\theta} = Q_0\bar{Y}_0 + Q_1\bar{Y}_1 + Q_2\bar{Y}_2$. Applying these results to Darwin's data we find that $\bar{Y}_0 = 20.97$.

To calculate $\bar{Y}_1$, we need r_i - the standardized residual, R_i, R and $\frac{R-\phi R_i}{(n-\phi)R}$ - the weight $\bar{Y}_1$ puts on the i-th observation. These values are given in Table 4.3 resulting in the value

$$\bar{Y}_1 = 24.65 .$$

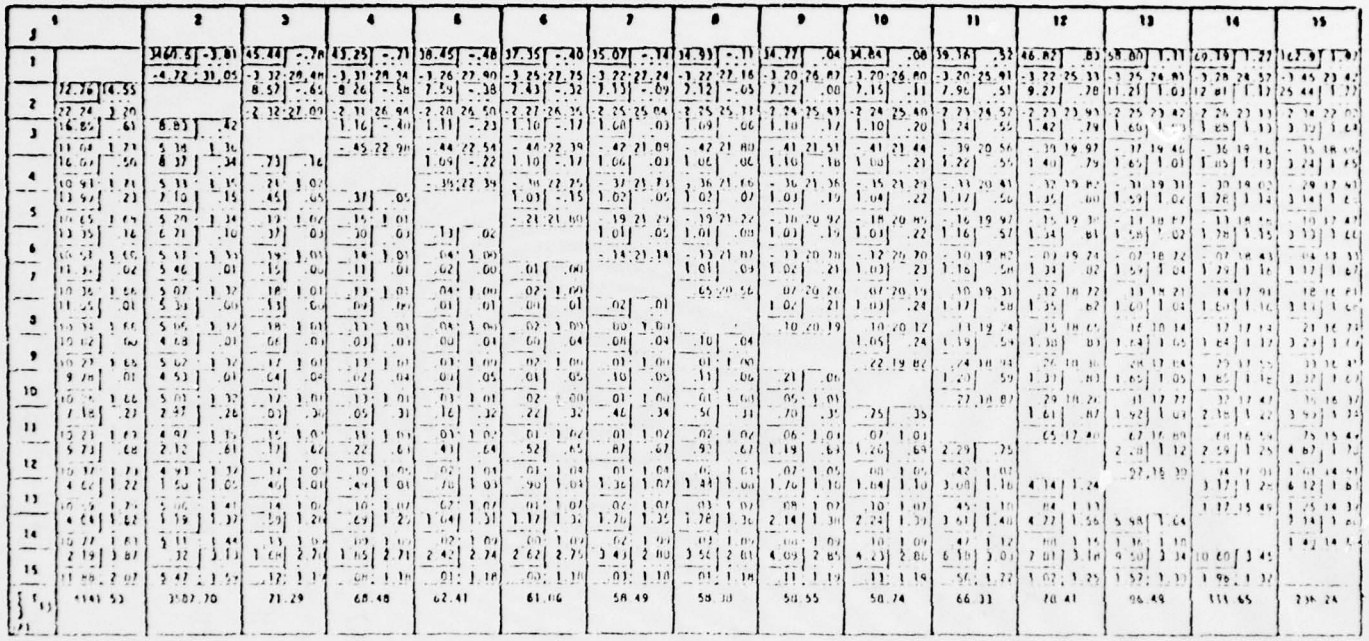
i	1	2	3	4	5	6	7	8
y_i	-67	-48	6	8	14	16	23	24
r_i	-2.33	-1.83	- .40	- .34	- .18	- .13	.05	.08
R_i	34.75	7.11	1.08	1.06	1.02	1.01	1.00	1.00
$R = 59.05$								
$\frac{R-\phi R_i}{(n-\phi)R}$	.03	.06	.07	.07	.07	.07	.07	.07

i	9	10	11	12	13	14	15
y_i	28	29	41	49	56	60	75
r_i	.19	.21	.53	.74	.93	1.04	1.43
R_i	1.02	1.02	1.16	1.34	1.58	1.77	3.13
$R = 59.05$							
$\frac{R-\phi R_i}{(n-\phi)R}$	.07	.07	.07	.07	.07	.07	.07

Table 3 Necessary information for calculating $\bar{Y}_1$

Note that R_i is very insensitive to r_i when $|r_i|$ is between zero and one and dramatically increases when $|r_i|$ becomes larger. For this sample R_1 is much larger than the others and results in a small weight for y_1.

The calculation of $\bar{Y}_2$ is more complicated; we need r_{ij}^i, r_{ij}^j, T_{ij}, T, and $\sum_{\substack{j=1 \\ j\neq i}}^{n} T_{ij}$. Table 4 gives all the relevant quantities. Since, by symmetry $T_{ij} = T_{ji}$, we use the upper



In the lower left triangle:

$(r^1_{ij}+r^j_{ij})^2$	$r^{1,j}_{ij}$
$r^{1,j}_{ij}$	A_{ij}

In the upper right triangle:

T_{ij}	r^1_{ij}
r^j_{ij}	$\bar{\theta}_{ij}$

$$r^1_{ij} = \frac{r_1 - \bar{\theta}_{ij}}{s_{ij}}$$

$$r^j_{ij} = \frac{r_j - \bar{\theta}_{ij}}{s_{ij}}$$

$$A_{ij} = 1 + \frac{1}{n-1}\left((r^1_{ij})^2 + (r^j_{ij})^2\right) - \frac{1}{n}(r^1_{ij}+r^j_{ij})^2$$

$$T_{ij} = A_{ij}^{\frac{1}{2}(n-1)}$$

$$T = \sum_{i<j} T_{ij}$$

Table 4 Relevant information for calculating $\dot{T}_2$

right half of the table to record r_{ij}^i, r_{ij}^j, T_{ij} and $\hat{\theta}_{ij}$, and the lower left half to record $(r_{ij}^i)^2$, $(r_{ij}^j)^2$, $(r_{ij}^i + r_{ij}^j)^2$ and $A_{ij} = 1 - \frac{\phi}{n-1}((r_{ij}^i)^2+(r_{ij}^j)^2 - \frac{\phi}{n}(r_{ij}^i + r_{ij}^j)^2)$. Their corresponding positions are explained at the bottom of Table 4.

For each i, $\sum_{\substack{j=1 \\ j\neq i}}^{n} T_{ij}$ is given in the last row and T is shown at the lower right corner. From these quantities, one can then calculate $\frac{T-\phi \sum_{\substack{j=1 \\ j\neq i}}^{n} T_{ij}}{(n-2\phi)T}$, the weight applied to the i-th observation (Table 5) and $\bar{Y}_2 = 30.74$.

i	1	2	3	4	5	6	7	8
y_i	-67	-48	6	8	14	16	23	24
$\frac{T-\phi \sum_{\substack{j=1 \\ j\neq i}}^{n} T_{ij}}{(n-2\phi)T}$	.007	.017	.075	.075	.075	.075	.075	.075

i	9	10	11	12	13	14	15
y_i	28	29	41	49	56	60	75
$\frac{T-\phi \sum_{\substack{j=i \\ j\neq i}}^{n} T_{ij}}{(n-2\phi)T}$	.075	.075	.075	.075	.075	.075	.073

Table 5 The weight applied to each observation in $\bar{Y}_2$

We again noticed that T_{ij} which is proportional to the inverse of a t ordinate at (r_{ij}^i, r_{ij}^j) is not sensitive to r_{ij}^i and r_{ij}^j when they are both between zero and one. But they increase at a very fast rate if either one gets large (over 1.5, say), and even faster if both are large. In Table 4 T_{15j} is larger than average for all j since the standardized residuals (r_{15j}^{15})'s are all larger than 1.5. T_{2j}'s are even larger but still not as dramatically large as T_{1j} for all j. And the dominating term is really T_{12} where both r_{12}^1 and r_{12}^2 are large. Under such circumstances, $\sum_{\substack{j=1 \\ j\neq 1}}^{n} T_{1j}$ and $\sum_{\substack{j=1 \\ j\neq 2}}^{n} T_{2j}$ are much larger than all other $\sum_{\substack{j=1 \\ j\neq i}}^{n} T_{ij}$ since they are the only two which include T_{12} in the summation. This results in the small weights for y_1 and y_2 in Table 5.

So far we have $\bar{Y}_0 = 20.93$, $\bar{Y}_1 = 24.65$, $\bar{Y}_2 = 30.74$ and furthermore we can calculate

$$Q_0 = \frac{1}{1+G \cdot 61.04 + G^2 \cdot 4720.35}$$

$$Q_1 = \frac{G \cdot 61.04}{1+G \cdot 61.04 + G^2 \cdot 4720.35}$$

$$Q_2 = \frac{G^2 \cdot 4720.35}{1+G \cdot 61.04 + G^2 \cdot 4720.35}$$

$$G = \frac{\alpha}{1-\alpha}\frac{1}{k}$$

The values taken by Q_o, Q_1, Q_2 depend on R, T and G. The larger G and T are, the larger Q_2 will be, the larger R is, the larger Q_1 will be. Since R and T are fixed once observations are obtained, and k is fixed at 5 for this example, the only thing we can change now is α (or G). Table 6 gives Q_o, Q_1, Q_2 and posterior mean $Q_o\bar{Y}_o + Q_1\bar{Y}_1 + Q_2\bar{Y}_2$ for different α (or G) values.

α	.01	.05	.10	.20
G	.002	.011	.022	.050
Q_o	.875	.462	.213	.063
Q_1	.108	.297	.289	.193
Q_2	.017	.242	.497	.744
$Q_o\bar{Y}_o+Q_1\bar{Y}_1+Q_2\bar{Y}_2$	21.50	24.41	26.89	28.95

Table 6 Q_o, Q_1, Q_2 and posterior means for different α (or G) values.

It is seen that different choices of α, and hence of G, greatly influence the value of Q_i and hence change the emphasis on $\bar{Y}_o$, $\bar{Y}_1$, and $\bar{Y}_2$, leading in this example to different results. When there are no bad values or very obvious bad values, the results are not particularly sensitive to the choice of α. In practice, it may be worthwhile to actually run the analysis on the computer and change the value of α (or G). Estimates which are

very sensitive to different choices of α (or G) indicate strongly, as in this example, the presence of bad values.

Empirical Study of the Dependency of the Posterior Mean on G

Recall we concluded that G is the only important parameter the results depend on when k is large. In the case of one bad value, we have shown that this is generally true for $k \geq 5$. For the general case of ℓ bad values, we believe this is still true but we probably will need a larger k.

Box and Tiao listed posterior means for Darwin's date with different choices of α and k; from these we can calculate G corresponding to each posterior mean, as shown in Table 7.

α \ k	5	6	7	8	9	10
.01	21.50 (.0020)	21.45 (.0017)	21.40 (.0014)	21.35 (.0013)	21.31 (.0011)	21.27 (.0010)
.02	22.21 (.0041)	22.11 (.0034)	22.00 (.0029)	21.89 (.0026)	21.79 (.0023)	21.71 (.0020)
.03	22.97 (.0062)	22.84 (.0052)	22.67 (.0044)	22.50 (.0039)	22.35 (.0034)	22.22 (.0031)
.04	23.71 (.0083)	23.56 (.0069)	23.36 (.0059)	23.14 (.0052)	22.94 (.0046)	22.76 (.0042)
.05	24.41 (.0105)	24.26 (.0088)	24.03 (.0075)	23.78 (.0066)	23.54 (.0058)	23.32 (.0053)
.06	25.03 (.0128)	24.90 (.0106)	24.66 (.0091)	24.39 (.0080)	24.13 (.0071)	23.87 (.0064)
.07	25.59 (.0151)	25.47 (.0125)	25.24 (.0108)	24.96 (.0094)	24.68 (.0084)	24.41 (.0075)
.08	26.08 (.0174)	25.99 (.0145)	25.77 (.0124)	25.49 (.0109)	25.21 (.0097)	24.92 (.0087)
.09	26.51 (.0198)	26.45 (.0165)	26.25 (.0141)	25.98 (.0124)	25.69 (.0110)	25.40 (.0099)
.10	26.89 (.0222)	26.86 (.0185)	26.67 (.0159)	26.42 (.0139)	26.13 (.0123)	25.85 (.0111)

Table 7 The posterior mean for Darwin's data with different α and k and the corresponding G given in the bracket under the posterior mean.

It is clear from the table that for fixed G, the posterior means are in general not too different, for example,

G = .0111	(α = .10, k = 10.0)	25.85
G = .0110	(α = .09, k = 9.0)	25.69
G = .0109	(α = .08, k = 8.0)	25.49
G = .0108	(α = .07, k = 7.0)	25.24
G = .0106	(α = .06, k = 6.0)	24.90
G = .0105	(α = .05, k = 5.0)	24.41

But there is a systematic change as k decreases indicating the dependency on k. It is also seen that such a change is smaller when k is larger. For this example, it seems that we need to have k at least as large as 7 for the results to be essentially dependent on G.

In general, this property would depend on the ratio $\frac{\text{number of bad values}}{\text{number of observations}}$. The larger this ratio is, the larger k must be for G to be the dominant parameter.

7. The 2^2 Factorial Design

It is much more difficult to obtain results in terms of weighting as soon as we deal with more complicated designs. As a simple example, consider the 2^2 factorial design with the linear model

$$\underset{\sim}{Y} = \underset{\sim\sim}{X}\theta + \underset{\sim}{\varepsilon}$$

where

$$\underset{\sim}{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & +1 & -1 \\ 1 & -1 & +1 \\ 1 & +1 & +1 \end{bmatrix} \qquad \underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} .$$

If we assume there is at most one bad value, we can obtain the posterior mean as follows:

Let $p = \frac{1}{k^2}$

$\hat{\theta}_{(0)}$ = posterior mean given there are no bad values

$$= \frac{1}{4} \begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ -y_1 + y_2 - y_3 + y_4 \\ -y_1 - y_2 + y_3 + y_4 \end{bmatrix}$$

$\hat{\theta}_{(1)}$ = the conditional posterior mean given that y_1 is a bad value

$$= \frac{1}{6p+2} \begin{bmatrix} 2py_1 + (p+1)y_2 + (p+1)y_3 + 2py_4 \\ -2py_1 + 2py_2 - (p+1)y_3 + (p+1)y_4 \\ -2py_1 - (p+1)y_2 + 2py_3 + (p+1)y_4 \end{bmatrix}$$

$\hat{\theta}_{(2)}$ = the conditional posterior mean given that y_2 is a bad value

$$= \frac{1}{6p+2} \begin{bmatrix} (p+1)y_1 + 2py_2 + 2py_3 + (p+1)y_4 \\ -2py_1 + 2py_2 - (p+1)y_3 + (p+1)y_4 \\ -(p+1)y_1 - 2py_2 + (p+1)y_3 + 2py_4 \end{bmatrix}$$

$\hat{\theta}_{(3)}$ = the conditional posterior mean given that y_3 is a bad value

$$= \frac{1}{6p+2} \begin{bmatrix} (p+1)y_1 + 2py_2 & +2py_3 & +(p+1)y_4 \\ -(p+1)y_1 & +(p+1)y_2 & -2py_3 + 2py_4 \\ -2py_1 & -(p+1)y_2 & +2py_3 +(p+1)y_4 \end{bmatrix}$$

$\hat{\theta}_{(4)}$ = the conditional posterior mean given that y_4 is a bad value

$$= \frac{1}{6p+2} \begin{bmatrix} 2py_1 & +(p+1)y_2 & +(p+1)y_3 & +2py_4 \\ -(p+1)y_1 & +(p+1)y_2 & -2py_3 & +2py_4 \\ -(p+1)y_1 & -2py_2 & +(p+1)y_3 & +2py_4 \end{bmatrix}$$

and from (5.10) we have

w_i = the posterior probability that the i-th observation is a bad value

$$= c\,\frac{\alpha}{1-\alpha}\,\frac{1}{k}\,\frac{8}{(4-\phi)^3-3(4-\phi)-2}\,\{1 + \phi r_i'(1-\tfrac{3}{4}\phi)r_i\}^{\frac{1}{2}}$$

$$= c\,\frac{\alpha}{1-\alpha}\,\frac{1}{k}\,\frac{8}{(4-\phi)^3-3(4-\phi)-2}\,\{1 + \phi(1-\tfrac{3}{4}\phi)r_i^2\}^{\frac{1}{2}} \quad .$$

The quantity w_0 is the posterior probability that there are no bad values and is equal to c. Then,

$$\text{the posterior mean} = \sum_{i=0}^{4} w_i \hat{\theta}_{(i)} \quad .$$

To see how this posterior mean weights each observation, we write the posterior mean as

$$\left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix}\begin{bmatrix} v_1^{-1} & 0 & 0 & 0 \\ 0 & v_2^{-1} & 0 & 0 \\ 0 & 0 & v_3^{-1} & 0 \\ 0 & 0 & 0 & v_4^{-1} \end{bmatrix}\begin{bmatrix} 1 & -1 & -1 \\ 1 & +1 & -1 \\ 1 & -1 & +1 \\ 1 & +1 & +1 \end{bmatrix}\right)^{-1}\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix}\begin{bmatrix} v_1^{-1} & 0 & 0 & 0 \\ 0 & v_2^{-1} & 0 & 0 \\ 0 & 0 & v_3^{-1} & 0 \\ 0 & 0 & 0 & v_4^{-1} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

then v_i^{-1} is the weight applied to the i-th observation in this Bayesian procedure.

By equating the above formula with the posterior mean, we can solve for v_i^{-1} and we have

$$v_1^{-1} \propto \frac{1}{(4-4w_1-w_0)p+4w_1+w_0}$$

$$v_2^{-1} \propto \frac{1}{(4-4w_2-w_0)p+4w_2+w_0}$$

$$v_3^{-1} \propto \frac{1}{(4-4w_3-w_0)p+4w_3+w_0}$$

$$v_4^{-1} \propto \frac{1}{(4-4w_4-w_0)p+4w_4+w_0} .$$

Since $p = \frac{1}{k^2}$ which is usually small, we can approximate p by zero and have approximately

$$v_1^{-1} \propto \frac{1}{4w_1+w_0} \qquad v_3^{-1} \propto \frac{1}{4w_3+w_0}$$

$$v_2^{-1} \propto \frac{1}{4w_2+w_0} \qquad v_4^{-1} \propto \frac{1}{4w_4+w_0}$$

The results are not surprising because we would expect that the weight given to the i-th observation would be small, if the posterior probability indicates that the i-th observation is bad.

For the general linear model where we do not have orthogonality, no general interpretative results of the kind possible for simple cases occur. However, of course, the Bayesian procedure described by Box and Tiao (1968) can still be carried through.

8. Summary

The Bayesian procedure proposed by Box and Tiao (1968) has been studied in further detail. We conclude that

(i) The posterior probability that a particular set of observations is discordant depends on the inverse of the multivariate-t ordinate of the standardized residuals corresponding to those observations.

(ii) In the location case, the posterior mean can be written as $Q_0\bar{Y}_0 + Q_1\bar{Y}_1 + \ldots + Q_\ell\bar{Y}_\ell$ if we assume at most ℓ bad values present. The quantity $\bar{Y}_i$ is the posterior mean given that there are i bad values and does not involve α. The quantity Q_i is the posterior probability of i bad values which involves both α and k.

(iii) If k is large enough so that ϕ is approximately one, $\bar{Y}_i$ does not depend heavily on k, and Q_i depends almost exclusively on $G = \frac{\alpha}{1-\alpha}\frac{1}{k}$. Thus, as an approximation, one can talk about this procedure in terms of one parameter G only. For this to be true in the location

case where there is only one bad value in a sample of size 10, simulation shows that k must be larger than five. As the number of bad values becomes larger relative to the sample size, a larger k value seems to be needed.

(iv) Bad values are given less weight if they are more evenly spread on both sides of the location parameter and more weight if they are more concentrated on either one side.

(v) For a 2^2 factorial design, it is shown one can also write the posterior mean in a weighted least square form. In this case, the weight for the i-th observation is approximately proportional to $\frac{1}{4w_i+w_o}$.

References

1. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), Robust Estimates of Location: Survey and Advances. Princeton Univ. Press.

2. Barnard, G. A. (1959), "Control charts and stochastic processes," JRSS, B, 21, p 239-257.

3. Beaton, A. F. and Tukey, J. W. (1974), "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," Technometrics, 16, p 147-185.

4. Box, G. E. P. (1978), "Robustness and Modelling," presented at ARO Workshop on Robustness.

5. Box, G. E. P. and Tiao, G. C. (1968), "A Bayesian approach to some outlier problems," Biometrika, 55, 1, p 119-129.

6. Dixon, W. J. (1953), "Processing data for outliers," Biometrics, 9, p 74-89.

7. Fisher, R. A. (1960), The Design of Experiments (7th edition). Edinburgh: Oliver and Boyd.

8. Huber, P. J. (1972), "Robust statistics: A review," Ann. Math. Statist., 43, p 1041-1067.

9. Stigler, S. M. (1977), "Do robust estimators work with real data?" The Annals of Statistics, 5, 6, P1055-1098.

10. Tukey, J. W. (1960), "A survey of sampling from contaminated distributions," Contributions to Probability and and Statistics: Volume Dedicated to Harold Hotelling. Stanford University Press.

11. Tukey, J. W. and McLaughlin, D. H. (1963), "Less vulnerable confidence and significance procedures for location based on a single sample (Trimming/Winsorization 1)," Sankhyā, Ser. A, 25, p 331-352.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 1998	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): FURTHER STUDY OF ROBUSTIFICATION VIA A BAYESIAN APPROACH.	5. TYPE OF REPORT & PERIOD COVERED: Summary Report, no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Gina Chen and George E. P. Box	8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 4 - Probability, Statistics, and Combinatorics
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709	12. REPORT DATE: September 1979
	13. NUMBER OF PAGES: 48
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Contaminated normal model, Bayesian procedure, M-estimator, weighting function, standardized residuals

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Bayesian outlier procedure discussed by Box and Tiao (1968) which uses the contaminated normal model is further explored in this report. For a simple location estimate suggested by their method, the weight given each observation is expressed explicitly in terms of standardized residuals so allowing comparison with M-estimators.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

221 200